学校编码：10384　　分类号_____密级_____

学号：23020070153685　　UDC _____

厦门大学

博士学位论文

恶意软件智能检测若干方法的研究及其应用

Research on Intelligent Malware Detection Methods and their Applications

叶艳芳

指导教师姓名：姜青山　教授

专业名称：人工智能基础

论文提交时间：2010 年 5 月

论文答辩日期：2010 年　月

学位授予日期：2010 年　月

答辩委员会主席：

评　阅　人：

2010 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：叶艳芳

2010 年 5 月 31 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：叶艳芳

2010 年 5 月 31 日

摘 要

随着网络应用的发展和安全形势的变化，互联网安全需求也随之有了新的变化和发展。恶意软件的爆发式增长和传播速度使得以客户端为战场的传统杀毒模式已经不能适应新的安全需求，为了与庞大而成熟的黑色产业链进行对抗，信息安全领域推出了相应的“云安全”计划。“云安全”是从“云计算”的概念中衍生而来，是信息安全界根据自身现状总结出来的概念，它融合了并行处理、网格计算、未知病毒行为判断等新兴技术和概念，通过网状的大量客户端对网络中软件行为的异常监测，获取互联网中病毒、木马等恶意软件的最新信息，传送到服务端进行自动分析和处理，再把病毒和木马的解决方案分发到每一个客户端。换句话说，反恶意软件与恶意软件的对抗战场由客户端转向了服务端。然而，在目前主流的“云安全”计划中，海量样本数据的分析普遍成为“云安全”实际应用的主要瓶颈。为了使“云安全”计划顺利实施，信息安全领域迫切需要新的技术和工具来智能地对大量的样本数据进行自动识别、分析和处理。研究数据挖掘的相关方法，并将其应用于信息安全领域是具有挑战而又有实际意义的课题。

本文从分析互联网安全现状和黑色产业链特点入手，针对“云安全”中恶意软件检测的难点和主要问题，研究了关联分类及分类集成学习的方法，将其应用于服务端恶意软件自动检测中，接着对聚类及聚类融合方法进行较深入的研究，将其应用于客户端恶意软件快速识别中，具有重要的理论意义和实际的应用价值。本研究的理论成果已成功转化到金山“云安全”的实际应用中。

本文的主要工作及贡献如下：

1. 深入分析了互联网安全的现状和黑色产业链的形成与发展，根据当前互联网安全形势和恶意软件检测的研究现状，阐述了“云安全”计划面临的难点和需要解决的主要问题；
2. 由于关联分类具有规则可解释性和分类准确的特点，本文对关联分类的方法进行了系统的研究，包括：分类关联规则的挖掘、分类关联规则的增量学习、规则的后处理方法、分类器的构造方法等，并成功应用于恶意软件检测中；

3. 基于不同的特征表达方法，研究了相应的分类算法，并针对异构集成个体产生数据缺失的情况进行深入研究，提出了一种适合缺失数据的分类集成学习方法；
4. 研究适合于恶意软件归类的文件特征表达方法，针对指令频度特征集提出了一种无需参数输入的层次与 *k-medoids* 混合的聚类方法 PFHK，针对高维稀疏的 N-Gram 指令序列特征集提出了一种特征加权的 *k-medoids* 聚类方法 WKM；
5. 在恶意软件归类中，经常需要引入专家知识，定义相应的约束条件，针对这种情况，提出了一种适合在约束条件下，对不同聚类算法进行融合的方法；
6. 将研究的分类、分类集成学习及聚类、聚类融合等相关方法成功应用于实际的恶意软件检测中。

关键词：恶意软件检测；关联分类；集成学习；聚类；聚类融合

Research on Intelligent Malware Detection Methods and their Applications

Abstract

The proliferation of malware has presented a serious threat to internet security. Currently, the most significant line of defense against malware is Anti-Virus (AV) software products which mainly use signature-based methods to recognize threats. However, malware writers quickly invented counter-measures against traditional signature-based methods employing obfuscation techniques such as polymorphism, metamorphism and packing to defeat efforts for analyzing the inner mechanisms of malware samples. In the area of internet security, the plan of "cloud security" is proposed which focuses on authenticating valid software from a white list, blocking invalid software from a black list, and analyzing any unknown software in a controlled manner (i.e., the gray list) at the server. However, along with the development of the malware industry, the number of file samples in the gray list that need to be analyzed on a daily basis is constantly increasing. The development trend of malware has motivated many research efforts on intelligent malware analysis, where data mining and machine learning techniques are used for malware detection. Such techniques have isolated successes in clustering and/or classifying particular sets of malware samples, but they have limitations that leave a large room for improvement and none of them have been applied in real applications.

In this dissertation, we investigate and develop advanced data mining methods for intelligent malware analysis. Different from earlier studies, our work is based on the large and real collection of file samples collected at Kingsoft anti-malware laboratory. We first provide an in-depth analysis of the development of malware; then systematically investigate and adapt associative classification and ensemble classification methods for malware detection; furthermore, we develop novel clustering algorithms to account for the characteristics of malware feature representations and propose a principled cluster ensemble framework for combining individual clustering solutions for malware categorization. The contributions in the dissertation have much important theoretical and practical significance and have been incorporated in real anti-malware applications.

The contributions of the dissertation are summarized as follows:

1. We provide an in-depth analysis of the development of malware ;
2. We provide a comprehensive feature extraction framework that combines four techniques for malware feature extraction;
3. We propose to use associative classifiers with post-processing techniques (including rule pruning, rule ordering and rule selection) to build interpretable classifiers for malware detection and develop an effective ensemble classification framework to combine heterogeneous base-level classifiers derived by different learning methods, using different feature representations on dynamic training sets.
4. To account for the characteristics of malware feature reprenestations, we propose a hybrid hierarchical clustering algorithm which combines the merits of hierarchical clustering and *k-medoids* algorithms and a weighted subspace *k-medoids* algorithm for malware categorization;
5. We propose a principled cluster ensemble framework for combining individual clustering solutions based on the consensus partition. The domain knowledge in the form of sample-level constraints can be naturally incorporated in the ensemble framework.
6. Finally, we evaluate and validate our proposed methods on the large and real daily sample collection from Kingsoft anti-malware laboratory. Promising experimental results demonstrate the effectiveness and efficiency of our proposed methods. Thus, they have been incorporated into Kingsoft anti-malware products.

Keywords: Malware detection; Associative classification; Classifier ensemble; Clustering; Clustering ensemble

目 录

Table of Contents